General Article

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(2) 198–218

DOI: 10.1177/2515245918771329
www.psychologicalscience.org/AMPPS

Statistical Rituals: The Replication Delusion and How We Got There

Gerd Gigerenzer
Harding Center for Risk Literacy, Max-Planck Institute for Human Development, Berlin, Germany

Abstract

The "replication crisis" has been attributed to misguided external incentives gamed by researchers (the *strategic-game hypothesis*). Here, I want to draw attention to a complementary internal factor, namely, researchers' widespread faith in a statistical ritual and associated delusions (the *statistical-ritual hypothesis*). The "null ritual," unknown in statistics proper, eliminates judgment precisely at points where statistical theories demand it. The crucial delusion is that the *p* value specifies the probability of a successful replication (i.e., 1 – *p*), which makes replication studies appear to be superfluous. A review of studies with 839 academic psychologists and 991 students shows that the replication delusion existed among 20% of the faculty teaching statistics in psychology, 39% of the professors and lecturers, and 66% of the students. Two further beliefs, the illusion of certainty (e.g., that statistical significance proves that an effect exists) and Bayesian wishful thinking (e.g., that the probability of the alternative hypothesis being true is 1 – *p*), also make successful replication appear to be certain or almost certain, respectively. In every study reviewed, the majority of researchers (56%–97%) exhibited one or more of these delusions. Psychology departments need to begin teaching statistical thinking, not rituals, and journal editors should no longer accept manuscripts that report results as "significant" or "not significant."

Keywords

replication, *p*-hacking, illusion of certainty, *p* value, null ritual

Received 6/28/17; Revision accepted 3/26/18

Every couple of weeks, the media proclaim the discovery of a new tumor marker that promises to improve personalized diagnosis or even treatment of cancer. As swift a pace as this seems, tumor research in fact produces many more discoveries. On average, four or five studies on cancer markers are published daily, almost all of them reporting at least one statistically significant prognostic marker (Ioannidis et al., 2014). Nonetheless, few of these results have been replicated and translated into clinical practice. When a team of 100 scientists at biotech company Amgen tried to replicate the findings of 53 "landmark" articles, they succeeded for only 6. Similarly, when the pharmaceutical company Bayer examined 67 projects on oncology, women's health, and cardiovascular medicine, they were able to replicate the results in only 14 cases (Mullard, 2011; Prinz, Schlange, & Asadullah, 2011).[1] In the United States alone, irreproducible preclinical research slowing down the discovery of life-saving therapies and cures has been estimated as costing $28 billion annually (Freedman, Cockburn, & Simcoe, 2015). The recently discovered fact that so many published results are apparently false alarms has been baptized the "replication crisis."

In the social sciences, replication studies were rarely published until recently, so the problem has lurked below the surface for many decades. An early analysis of 362 articles in psychology found no attempted single replication study for any of them (Sterling, 1959), a subsequent analysis found replications for fewer than 1% of more than 1,000 articles (Bozarth & Roberts, 1972), and a 2012 analysis found replications for 1% of original articles (Makel, Plucker, & Hegarty, 2012). By

Corresponding Author:
Gerd Gigerenzer, Director, Harding Center for Risk Literacy, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany
E-mail: gigerenzer@mpib-berlin.mpg.de

2010, the crisis began to shake up the social sciences, after it was reported that highly publicized results could not be replicated (e.g., Open Science Collaboration, 2015; Pashler, Coburn, & Harris, 2012; Scheibehenne, Greifeneder, & Todd, 2010). At the beginning of the 21st century, one of the most cited claims in the social and biomedical sciences was that most scientific results are false (Ioannidis, 2005; Szucs & Ioannidis, 2017). Is science on its last legs? Its detractors certainly appear keen to cash in on the crisis. On March 29, 2017, the news Web site Breitbart.com headlined a claim by Wharton School professor Scott Armstrong that "fewer than 1 percent of papers in scientific journals follow scientific method" (Bokhari, 2017), which was widely spread across other conservative opinion and news Web sites. For politicians and citizens skeptical of science, such messages provide a perfect reason to justify cuts in funding for research that conflicts with their beliefs and values.

Only a Matter of Incentives?

The replication crisis has led to heated debates on how much waste is produced, in what fields this occurs, and what exactly counts as a replication in the first place (e.g., Pashler & Wagenmakers, 2012; Stroebe & Strack, 2014). In this article, I do not deal with these highly debated issues but rather consider the question of how we got here. Little is known about the causes of the problem. Some people have suggested that truth simply wears off, as when the efficacy of antidepressants plummeted drastically from study to study. In the 1930s, Joseph B. Rhine concluded that the extrasensory perception of his students declined over the years. And one cognitive psychologist, sincerely puzzled by how a much-publicized effect that he discovered could have faded away, invoked the concept of "cosmic habituation," a natural law postulating that once a claim for an effect is published, the world habituates to it so that it begins shrinking away (Lehrer, 2010; see also Schooler, 2011).

More seriously, the general approach has been to explain the replication crisis by blaming wrong economic and reputational incentives in the academic career system. Specifically, the blame has been laid on the "publish or perish" culture, which values quantity over quality, and on the practice of evaluating researchers by impact measures such as h-index and, more recently, "altmetrics" such as the number of tweets and mentions in blogs (Colquhoun, 2014). Richard Horton (2016), editor-in-chief of *The Lancet*, complained: "No-one is incentivized to be right" (p. 1380). In this view, powerful incentives for career advancement actively encourage, reward, and propagate poor research methods and abuse of statistics (e.g., Smaldino & McElreath, 2016). The various symptoms of wrong incentives include hyped-up press releases and misleading abstracts that claim discoveries even when they are not supported by the data; publication bias, that is, the reluctance of journals and authors to publish negative results; lack of willingness to share data; financial conflicts of interest (particularly in clinical research; Schünemann, Ghersi, Kreis, Antes, & Bousquet, 2011); and commodification and privatization of research (Mirowski, 2011). Following Smaldino and McElreath (2016), I refer to these explanations as the *strategic-game hypothesis*, according to which science is considered a game that scientists play strategically to maximize their chances of acquiring publications and other trophies (Bakker, van Dijk, & Wicherts, 2012). To counter gaming, preregistration and data sharing are now encouraged or required by several journals, and radical changes to the reward system have been proposed. One drastic proposal to revamp the current system is that only publications whose findings are replicated, not publication per se, should count in the future and that large grants should count not positively but negatively, unless the recipient delivers proportionally high-quality science (Ioannidis, 2014).

In this article, I present an analysis that goes beyond the important role of external incentives, be they economic or reputational. I discuss the hypothesis that the replication crisis is also fueled by an internal factor: the replacement of good scientific practice by a statistical ritual that researchers perform not simply on the grounds of opportunism or incentives but because they have internalized the ritual and genuinely believe in it. I refer to this hypothesis as the *statistical-ritual hypothesis*. As is the case with many social rituals, devotion entails delusions, which in the present case block judgment about how to conduct good research, that is, inhibit researchers' common sense. I use the term *common sense* because the delusions in question here do not concern sophisticated statistical technicalities but instead concern the very basics. If the cause were merely strategic behavior, common sense would not likely be sacrificed.

This article comprises two main parts. In the first, I show how textbook writers created the null ritual as an apparently objective procedure to distinguish a true cause from mere chance. This procedure is aimed at yes/no conclusions from single studies, neglects replication and other principles of good scientific practice, and has become dominant in precisely those sciences involved in the replication crisis. It is worth noting that editors in the natural sciences, which do not practice the ritual, generally endorse replication and do not separate it from original research (Madden, Easley, & Dunn, 1995). In the second part of this article, I review studies on four phenomena that are implied by the

statistical-ritual hypothesis but are not explained (or, in one case, are only partially explained) by the strategic-game hypothesis. In general, the strategic-game hypothesis implies that researchers play the game without necessarily being in the grip of systematic delusions or acting against their own interests. The statistical-ritual hypothesis, in contrast, implies that researchers engage in delusions about the meaning of the null ritual, and above all about its sacred number, the *p* value. Otherwise, they would realize that the *p* value does not answer their research questions and would abandon its pursuit. The first phenomenon is the replication delusion, which makes replication appear virtually certain and further studies superfluous. The second concerns the illusion of certainty, and the third Bayesian wishful thinking; both of these lead to the same conclusions about replication. A review of studies with 839 academic psychologists shows that the majority believed in one or more of these three fallacies. Finally, there is the phenomenon of low statistical power. According to the strategic-game hypothesis, researchers should be careful to design experiments that have a good chance of detecting an effect and thus lead to the desired result, statistical significance. The statistical-ritual hypothesis, however, implies that researchers pay little attention to statistical power because it is not part of the null ritual. Consistent with the latter prediction, studies of the psychological literature show that the median statistical power for detecting a medium effect size is generally low and has not improved in the past 50 years. The statistical-ritual hypothesis also accounts for why the specific incentives that strategic behavior exploits were set in the first place.

The Idol of Automatic Inference and the Elimination of Judgment

Statistical methods are not simply applied to a discipline but can transform it entirely. Consider medicine. For centuries, its lifeblood was physicians' "medical tact" and the search for deterministic causes. All that changed when, in the second half of the 20th century, probabilities from randomized trials and *p* values replaced causes with chances. Or think of parapsychology, in which statistical tests became common half a century earlier. Once the study of unique messages from the dear departed, extrasensory perception is now the study of repetitive card guessing; marvels have been replaced with statistical significance (Gigerenzer et al., 1989).

Psychology has also been transformed by the probabilistic revolution (Gigerenzer, 1987). For the present topic, two aspects of this remarkable event are of relevance. First, as in medicine, the probabilistic revolution forged an unexpected marriage between two previously antagonistic tribes: experimenters and statisticians. In fact, statistical inference became the hallmark of experimentation, and experiments without statistical inference were soon virtually unthinkable. This change was so profound that quite a few social scientists today would be surprised to learn that Isaac Newton, for instance, used no statistical tests in his experiments. You might object that Newton was not familiar with statistical inference, but in fact, he was: In his role as the master of the London Royal Mint, he used statistical tests to make sure that the amount of gold in the coins was neither too small nor too large (Stigler, 1999). The Trial of the Pyx involved random samples of coins, a null hypothesis to be tested (that the tested coin conformed to the standard coin), a two-sided alternative hypothesis, and a test statistic. In Newton's view, statistical tests were useful for quality control, but not for science. Similarly, 19th-century medicine saw professional rivalry between experimenters, who looked for causes, and statisticians, who looked for chances. For instance, physiologist Claude Bernard, one of the first to suggest blind experiments, opposed the use of averages or proportions as unscientific. For him, averages were no substitutes for a complete investigation of the conditions that cause variability (Gigerenzer et al., 1989, pp. 126–130). The British biologist and statistician Sir Ronald A. Fisher (1890–1962) was highly influential in ending this antagonism and forging a marriage between experimenters and statisticians.

Second, and most relevant for this article, psychologists reinterpreted this marriage in their own way. Early textbook writers struggled to create a supposedly objective method of statistical inference that would distinguish a cause from a chance in a mechanical way, eliminating judgment. The result was a shotgun wedding between some of Fisher's ideas and those of his intellectual opponents, the Polish statistician Jerzy Neyman (1894–1981) and the British statistician Egon S. Pearson (1895–1980). The essence of this hybrid theory (Gigerenzer, 1993) is the null ritual. The null ritual does not exist in statistics proper, although researchers have been made to believe so.

The inference revolution

The term *inference revolution* refers to a change in scientific practice that happened in psychology between 1940 and 1955 and subsequently in other social sciences and biomedical research (Gigerenzer & Murray, 1987). The inference from sample to population came to be considered the sine qua non of good research, and statistical significance came to be considered the means of distinguishing between true cause and mere

chance. Common scientific standards such as minimizing measurement error, conducting double-blind experiments, replicating experiments, providing detailed descriptive statistics, and formulating bold hypotheses in the first place were pushed into the background (Danziger, 1990; Gigerenzer, 1993). The focus on inference from sample to population is remarkable or even perplexing given that in most psychological experiments, researchers do not draw random samples from a population or define a population in the first place. Thus, the assumptions for the procedure are not met in most cases, and one does not know the population to which an inference such as $\mu_1 \neq \mu_2$ actually refers.

The term *revolution* underscores the dramatic impact of this change. In the earlier experimental tradition, the unit of analysis was the individual, who was well trained and often held a Ph.D. After the revolution, the unit became the aggregate, often consisting of minors or undergraduates. In the United States, this change was already under way during the 1930s, when psychologists tried to promote their field as socially relevant, and educational administrators were the key source of funding. For administrators, the relevant question was whether a new teaching method led to better performance on average; to that end, psychologists created a new unit, the treatment group (Danziger, 1987). But when a mean difference between two groups was observed, they had to use their judgment about whether it was "real." When psychologists learned about Fisher's method of statistical inference (see the section on the null ritual), it seemed like a dream come true: an apparently objective method to tell a cause from a chance.

Accordingly, early adopters of significance testing were concentrated in educational psychology and parapsychology, while the core scientific disciplines, such as perceptual psychology, resisted the movement for years and continued to focus on individuals. It is instructive that the situation in Germany was different (Danziger, 1990). German professors did not feel pressured to prove the usefulness of their research for education but instead considered themselves scientists responsible for unraveling the laws of the individual mind. Accordingly, the inference revolution happened much later in German psychology, in its post–World War II assimilation into U.S. psychology. The changes in research practice spawned by the inference revolution were so fundamental that psychologists trained in its wake can hardly imagine that research could entail anything other than analyzing means of aggregates for statistical significance.

What did psychologists do before the inference revolution?

A typical article in the *Journal of Experimental Psychology* around 1925 would report single-case data in detail, using means, standard deviations, correlations, and various descriptive statistics tailored to the problem. The participants were trained staff members with Ph.D.s or graduate students (Danziger, 1990). In the tradition of Wilhelm Wundt's Leipzig laboratory, the researcher who ran the experiment was typically a technician, which is why it could happen that the participant, not the researcher, published the article. When a larger number of individuals were studied, as in Jean Piaget's work on the development of cognition, results were typically presented individually rather than as aggregates—you would not have caught Piaget calculating a *t* test. The overall picture is that past researchers knew their raw data well, reported descriptive statistics in detail, and had a comparatively flexible attitude toward the issue of statistical inference from sample to population, which was considered to be incidental rather than central.

When treatment groups began to be used in the applied research of the 1930s in the United States, the term *significant* was already widely in use, but the evaluation of whether two means differed was based on judgment—taking the error, the size of the effect, and previous studies into account—rather than on a mechanical rule, except for the use of critical ratios (the ratio of the obtained difference to its standard deviation; Gigerenzer & Murray, 1987, chap. 1). This practice led to virtually all of the classical discoveries in psychology. Without calculating *p* values or Bayes factors, Wolfgang Köhler developed the Gestalt laws of perception, Ivan P. Pavlov the principles of classical conditioning, B. F. Skinner those of operant conditioning, George Miller his magical number seven plus or minus two, and Herbert A. Simon his Nobel Prize–winning work on bounded rationality.

The null ritual

In the 1920s, Fisher had forged the marriage between experiments and inferential statistics in agriculture and genetics, an event that went largely unnoticed by psychologists (for a discussion of Fisher's antecedents, see Gigerenzer et al., 1989, pp. 70–90). Through statisticians George W. Snedecor at Iowa State College and Harold Hotelling at Columbia University, among others, Fisher's theory of null-hypothesis testing soon spread in the United States. By 1961, Snedecor's (1937) *Statistical Methods* became the most cited book according to the Science Citation Index (Gigerenzer & Murray, 1987, p. 21). Psychologists began to cleanse Fisher's message of its agricultural odor—the effect of manure, soil fertility, and the weight of pigs—as well as of its mathematical sophistication, and wrote a new genre of textbooks. The most widely read of these was probably *Fundamental Statistics in Psychology and Education*, which was published in 1942 and written by J. P. Guilford, a

psychologist at the University of Southern California and later president of the American Psychological Association. Like Guilford, authors of best-selling statistical textbooks in psychology have typically been nonstatisticians.

Soon things got complicated, however. The statistical theory of Neyman and Pearson also became known, particularly after World War II. The essential differences relevant for our purpose are threefold: First, whereas Fisher proposed testing a single specified hypothesis, the null, against an unspecified alternative, Neyman and Pearson questioned this logic and called for testing against a second specified hypothesis. Second, with only a null and no specified alternative, Fisher had no measure of statistical power. Moreover, he believed that calculating power made sense in quality control but not in science—according to him, there is no place for cost-benefit trade-offs such as between power and alpha when one is seeking the truth. (Alpha is the probability that the null hypothesis is rejected if it is true, and beta, calculated as 1 – power, is the probability that the alternative hypothesis is rejected if it is true. Alpha and beta are also called Type I and Type II error rates, respectively.) Neyman and Pearson, in contrast, required power and alpha to be balanced and set before the experiment so that the probability of Type I and Type II errors would be known. Third, Fisher interpreted a significant effect in terms of subjective confidence in the result, whereas Neyman (albeit not Pearson) interpreted significance in strictly behavioristic terms as a decision, not a belief. For example, in quality control, misses matter, and a significant result can lead to the decision to stop production and look for a possible error, even when one believes that most likely nothing is wrong. Neyman regarded his theory as an improvement of null-hypothesis testing, but Fisher disagreed. The conceptual differences between these two systems of statistical inference were amplified by a fierce personal debate. Fisher branded Neyman's theory as "childish" and "horrifying [for] the intellectual freedom of the west," while Neyman countered that some of Fisher's tests were "worse than useless" because their power was smaller than their alpha level (see Gigerenzer et al., 1989, chap, 3). How should textbook writers have coped with the fundamental disagreement between these two camps? The obvious solution would have been to present both approaches and discuss the situations in which each might be more appropriate.

But such a toolbox approach would have required relinquishing dearly coveted objectivity and opened the door to researchers' judgment. Although a few textbooks did take this approach and taught both theories (e.g., R. L. Anderson & Bancroft, 1952), the great majority fused the two antagonistic theories into a hybrid theory, of which neither Fisher nor Neyman and Pearson would have approved. In addition to the idol of automatic inference, another decisive factor appears to have been the commercialization of textbooks, accompanied by publishers' requests for single-recipe cookbooks instead of a toolbox (Gigerenzer, 2004, pp. 587–588). To this end, virtually all textbooks presented the hybrid theory anonymously, without mentioning that its concepts stem from different theories, detailing the conflicting ideas of the theories' authors, or even disclosing their identities. For instance, although the 1965 edition of Guilford's best-selling *Fundamental Statistics in Psychology and Education* cites some 100 authors in its index, the names of Neyman and Pearson are left out.

The essence of this hybrid theory is the null ritual (Gigerenzer, 2004):

1. Set up a null hypothesis of "no mean difference" or "zero correlation." Do not specify the predictions of your own research hypothesis.
2. Use 5% as a convention for rejecting the null hypothesis. If the test is significant, accept your research hypothesis. Report the test result as $p < .05$, $p < .01$, or $p < .001$, whichever level is met by the obtained p value.
3. Always perform this procedure.

The null ritual does not exist in statistics proper. This point is not always understood; even its critics sometimes confuse it with Fisher's theory of null-hypothesis testing and call it "null-hypothesis significance testing." In fact, the ritual is an incoherent mishmash of ideas from Fisher on the one hand and Neyman and Pearson on the other, spiked with a characteristically novel contribution: the elimination of researchers' judgment.

Elimination of judgment

Consider Step 1 of the null ritual. To specify the null hypothesis but not an alternative hypothesis follows Fisher's logic but violates that of Neyman and Pearson, which necessitates specifying both (this is one of the reasons why they thought of their theory as an improvement over Fisher's). Yet Step 1 follows Fisher only to a point, because Fisher at least thought that judgment would be necessary for choosing a proper null hypothesis. As Fisher emphasized, his intent was to test whether a hypothesis should be nullified, and he did not mean to imply that this hypothesis postulates a nil difference (Fisher, 1955, 1956). In his approach, researchers should use their judgment to select a proper null hypothesis, which could be a nonzero difference or correlation. In the hands of the textbook writers in

psychology, however, *null* grew to mean "no difference," period. No judgment was required.

Step 2 blatantly contradicts Fisher, who in the 1950s advised researchers against using 5% in a mechanical way:

> No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. (Fisher, 1956, p. 42)[2]

Using a conventional 5% level of significance is also alien to Neyman and Pearson, who in their earliest writings already emphasized that their tests should be "used with discretion and understanding," depending on the context (Neyman & Pearson, 1933, p. 296). Understanding includes making a judgment about the balance between Type I and Type II errors; this judgment was also omitted in the null ritual.

The practice of rounding up p values to the next convenient "significance level" ($p < .05$, $p < .01$, or $p < .001$) is supported neither by Fisher nor by Neyman and Pearson. In their approaches, a level of significance is set before the experiment, not calculated post hoc from data, and p values calculated from data should be reported as exact values (e.g., $p = .004$, not $p < .01$) so that they are not confused with significance levels. Step 2 is also inconsistent with Fisher's (1955) argument against making binary reject/not-reject decisions, and appears to follow Neyman and Pearson. Yet it does so only partially and neglects the rest of their theory, which requires two precise hypotheses and a judgment about the balance between alpha and beta for determining the decision criterion. In the null ritual, there is no concern with beta and consequently no concern with the statistical power of a test, which is the complement of beta.

Finally, Step 3 embodies the idol of automatic inference that does not require a judgment about the validity of the assumptions underlying each statistical test. This is alien to both camps, and to statistical science in general. Fisher (1955, 1956) asserted that constructive imagination and much experience are prerequisites of good statistics, such as for deciding which null hypotheses are worth testing and which test statistic to choose. Neyman and Pearson emphasized that the statistical part of inference has to be supplemented by a subjective part. In Pearson's (1962) words:

> We left in our mathematical model a gap for the exercise of a more intuitive process of personal judgment in such matters—to use our terminology—as the choice of the most likely class of admissible hypotheses, the appropriate significance level, the magnitude of worthwhile effects and the balance of utilities. (pp. 395–396)

Throughout their heated personal debates, each side accused the other of advocating mechanical statistical inference. At least here, they saw eye to eye: Statistical inference should not be automatic. Yet it has become not only automatic but also what I call ritualistic, embodying social emotions, including fear, wishful thinking, and delusions, along with mechanical repetition.

I use the term *ritual* for the core of the hybrid theory to highlight its similarity to social rites (Dulaney & Fiske, 1994). A ritual is a collective or solemn ceremony consisting of actions performed in a prescribed order. It typically includes the following elements:

- sacred numbers or colors,
- repetition of the same action,
- fear about being punished when one stops performing these actions, and
- wishful thinking and delusions.

The null ritual contains all these features: a fixation on the 5% number (or on colors, as in functional MRI images), repetitive behavior resembling compulsive hand washing, fear of sanctions by editors or advisors, and delusions about the meaning of the p value (as I discuss in the next section). It spread fast. Before 1940, few articles reported t tests, analyses of variance, or other significance tests. By 1955 and 1956, more than 80% of articles in four leading journals were doing so, and almost all reported a significant effect (Sterling, 1959). Today, the figure is close to 100%.

Systematic Delusions About Replication

In this section, I put forward the hypothesis that the null ritual is key to understanding the replication crisis in the social and biomedical sciences. As I mentioned earlier, one explanation of this crisis has been that researchers are given the wrong incentives. These do, of course, exist, creating the fear of being punished for not conforming, the third aspect of a social ritual. Yet incentives are only part of the explanation, as is demonstrated by a rare case in which an editor actually eliminated them. When Geoffrey Loftus became editor-elect of *Memory & Cognition*, he made it clear in his introductory editorial that he did not want authors to submit manuscripts with routine calculations of p values, but instead wanted adequate figures with descriptive statistics and confidence intervals (Loftus, 1993). During his editorship, I asked him how his campaign was

going. Loftus bitterly complained that many researchers stubbornly refused the opportunity, experienced deep anxiety at the prospect of abandoning *p* values, and insisted on their *p* values and yes/no significance decisions (Gigerenzer, 2004, pp. 598–599). Similarly, after the Task Force for Statistical Inference's recommendations calling for various alternatives to the null ritual were incorporated in the fifth edition of the American Psychological Association's (2001) publication manual, a study found little change in researchers' behavior (Hoekstra, Finch, Kiers, & Johnson, 2006).

My hypothesis is that, beyond incentives, the key issue is researchers' internalized belief in the ritual. If researchers only opportunistically adapt their behavior to the incentives in order to get published and promoted, then their common sense with regard to statistical thinking should remain intact. If, however, researchers have internalized the ritual and believe in it, conflicting common sense should be repressed. The most basic and crucial test of the statistical-ritual hypothesis is whether researchers actually understand the desired product: a significant *p* value.

Probability of replication = 1 – p

A *p* value is a statement about the probability of data, assuming the null hypothesis is true. More precisely, the term *data* refers to a test statistic (a statistical summary of data, such as a *t* statistic), and the term *null hypothesis* refers to a statistical model. For instance, a result with a *p* value of .05 means that if the null hypothesis—including all assumptions made by the underlying model—is true, the probability of obtaining such a result or a more extreme one is 5%. The *p* value does not tell us much else. Specifically, a *p* value of .05 does not imply that the probability that the result can be replicated is 95%.

Consider a simple example: A researcher designs an experiment with 50% power to detect an effect of a medium size (50% power or below is a typical figure, as I discuss later) and obtains a significant difference between means, $p < .05$. Does this imply that one can expect to find a significant result in 95% (or more) of exact replications of this experiment? No, and the reason why not can be easily understood. If the alternative hypothesis is true, then the probability of getting another significant result equals the statistical power of the test, that is, 50%. If, however, the null hypothesis is true, the probability of getting another significant result would still be only 5%. Or consider an even simpler illustration: A die, which could be fair or loaded, is thrown twice and shows a "six" both times, which results in a *p* value of .03 (1/36) under the null hypothesis of a fair die. Yet this does not imply that one can expect two sixes in 97% of all further throws. In general, the chance of replicating a finding depends on many factors (e.g., Cumming, 2008, 2014; Greenwald, Gonzalez, Harris, & Guthrie, 1996), most of which the researcher cannot know for sure, such as whether the null or the alternative hypothesis is true. The belief that an obtained *p* value implies a probability of $1 - p$ that an exact replication of the same experiment would lead to a significant result is known as the *replication delusion* or *replication fallacy* (Rosenthal, 1993).

Note that this delusion is easy to see through, and experienced researchers should not hold this belief even if they follow the null ritual for opportunistic motives. In contrast, if they believe in the ritual, they are likely to exhibit the replication delusion because it provides a justification for performing the ritual, even if it amounts to wishful thinking. Thus, the empirical question is, do experienced researchers actually suffer from the replication delusion?

The first study to answer this question appears to have been conducted by Oakes (1986). He asked 70 British lecturers, research fellows, and postgraduate students with at least 2 years of research experience whether the following statement is true or false, that is, whether it logically follows from a significant result ($p = .01$):

> You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions. (p. 80)

Sixty percent of the psychologists answered "true." Table 1 shows that this replication delusion is also endemic beyond academic psychologists in the United Kingdom. Aggregating across all the studies listed in Table 1 leads to the following general picture: Among psychologists who taught statistics or methodology, 20% (23 of 115) believed in the replication delusion; among academic psychologists whose special field was not statistics, this number was larger, 39% (282 of 724); and among students, it was even larger, 66% (659 of 991).

The existence of the replication delusion is consistent with the hypothesis that a substantial number of researchers follow the null ritual not simply for strategic reasons but also because they believe in the ritual and its associated delusions. According to the replication delusion, given $p = .01$, a study's results can be replicated in 99% of all trials, which means that replication studies are superfluous.

The limits of the present analysis are the small number of studies that have been conducted and the low response rates reported in some of the studies. Badenes-Ribera, Frias-Navarro, Monterde-i-Bort, and Pascual-Soler (2015) sent their survey to 4,066 academic psychologists

Table 1. Studies on the Replication Delusion

Study	Description of group	Country	*N*	Statistic tested	Respondents exhibiting the replication delusion (%)
		Professional samples			
Oakes (1986)	Academic psychologists	United Kingdom	70	$p = .01$	60
Haller & Krauss (2002)	Statistics teachers in psychology	Germany	30	$p = .01$	37
Haller & Krauss (2002)	Professors of psychology	Germany	39	$p = .01$	49
Badenes-Ribera, Frias-Navarro, Monterde-i-Bort, & Pascual-Soler (2015)	Academic psychologists: personality, evaluation, psychological treatments	Spain	98	$p = .001$	35
Badenes-Ribera et al. (2015)	Academic psychologists: methodology	Spain	67	$p = .001$	16
Badenes-Ribera et al. (2015)	Academic psychologists: basic psychology	Spain	56	$p = .001$	36
Badenes-Ribera et al. (2015)	Academic psychologists: social psychology	Spain	74	$p = .001$	39
Badenes-Ribera et al. (2015)	Academic psychologists: psychobiology	Spain	29	$p = .001$	28
Badenes-Ribera et al. (2015)	Academic psychologists: developmental and educational psychology	Spain	94	$p = .001$	46
Badenes-Ribera, Frias-Navarro, Iotti, Bonilla-Campos, & Longobardi (2016)	Academic psychologists: methodology	Italy, Chile	18	$p = .001$	6
Badenes-Ribera et al. (2016)	Academic psychologists: other areas	Italy, Chile	146	$p = .001$	13
Hoekstra, Morey, Rouder, & Wagenmakers (2014)	Researchers in psychology (Ph.D. students and faculty)	Netherlands	118	95% CI	58
		Student samples			
Haller & Krauss (2002)	German psychology students who had passed two statistics courses	Germany	44	$p = .01$	41
Hoekstra et al. (2014)	First-year psychology students who had not taken an inferential-statistics course	Netherlands	442	95% CI	66
Hoekstra et al. (2014)	Master's psychology students who had taken an inferential-statistics course	Netherlands	34	95% CI	79
Garcia-Pérez & Alcalá-Quintana (2016)	First-year psychology students who had not taken an inferential-statistics course	Spain	313	95% CI	71
Garcia-Pérez & Alcalá-Quintana (2016)	Master's psychology students who had taken an inferential-statistics course[a]	Spain	158	95% CI	63

Note: For the studies in which the replication delusion was tested with respect to *p* values, the numbers in the last column indicate the percentage of respondents who erroneously believed that $p = .01$ or $p = .001$ implies that the probability of a significant result in a replication study is .99 or .999, respectively. Haller and Krauss (2002) used the same question used by Oakes (1986; see the text). The problem posed by Badenes-Ribera et al. (2015, 2016) read: "Let's suppose that a research article indicates a value of p = 0.001 in the results section (alpha = 0.05). Mark which of the following statements are true (T) or false (F)" (Badenes-Ribera et al., 2015, p. 291). The statement expressing the replication fallacy read: "A later replication would have a probability of 0.999 (1 – 0.001) of being significant" (p. 291). For the studies in which the replication delusion was tested with respect to confidence intervals (CIs), the numbers in the last column indicate the percentage of respondents who wrongly believed that a 95% CI ranging from *x* to *y* would imply that if the experiment were repeated over and over, the true mean would fall between *x* and *y* 95% of the time.

[a]A subsample of 88 of these students were considered by the authors to have provided "informed responses" (p. 10). Of that subsample, 72% exhibited the replication delusion.

in Spain and had a response rate of 10.3%; the study finding the lowest rate of the replication delusion, conducted by Badenes-Ribera, Frias-Navarro, Iotti, Bonilla-Campos, and Longobardi (2016), also had the lowest response rate, 7% (164 out of 2,321 academic psychologists in Italy and Chile); response rates for the other studies were not reported. Thus, the results may be subject to a selection bias: Assuming that individuals with better training in statistics were more likely to respond, the numbers in Table 1 are probably underestimates of the true frequency of the replication delusion.

A study with members of the Mathematical Psychology Group and the American Psychological Association (not included in Table 1 because the survey asked different kinds of questions) also found that most of them trusted in small samples and had high expectations about the replicability of significant results (Tversky & Kahneman, 1971). A glance into textbooks and editorials reveals that the delusion was already promoted as early as the 1950s. For instance, in her textbook *Differential Psychology*, Anastasi (1958) wrote: "The question of statistical significance refers primarily to the extent to which similar results would be expected if an investigation were to be repeated" (p. 9). In his *Introduction to Statistics for Psychology and Education*, Nunnally (1975) stated: "If the statistical significance is at the 0.05 level . . . the investigator can be confident with odds of 95 out of 100 that the observed difference will hold up in future investigations" (p. 195). Similarly, former editor of the *Journal of Experimental Psychology* A. W. Melton (1962) explained that he took the level of significance as a measure of the "confidence that the results of the experiment would be repeatable under the conditions described" (p. 553).

The illusion of certainty and Bayesian wishful thinking

As I have mentioned, a p value is a statement about the probability of a statistical summary of data, assuming that the null hypothesis is true. It delivers probability, not certainty. It does not tell us the probability that a hypothesis—whether the null or the alternative—is true; it is not a Bayesian posterior probability. I refer to the belief that statistical significance delivers certainty as the *illusion of certainty* and to the belief that p is the probability that the null hypothesis is true or that $1 - p$ is the probability that the alternative hypothesis is true as *Bayesian wishful thinking* (also known as *inverse probability error*). After a few hours of statistical training at a major university, any person of average intelligence should understand that these beliefs are incorrect. By contrast, if the statistical-ritual hypothesis is true, researchers' thinking should be partly blocked and they should endorse these beliefs about the importance of significant results.

Table 2 reviews the relevant studies that have been conducted. In the British study mentioned earlier, Oakes (1986, p. 80) asked academic psychologists what a significant result ($p = .01$) means:

> Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant (t = 2.7, df = 18, p = .01). Please mark each of the statements below as "true" or "false." "False" means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.
>
> (1) You have absolutely disproved the null hypothesis (i.e., there is no difference between the population means).
> (2) You have found the probability of the null hypothesis being true.
> (3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
> (4) You can deduce the probability of the experimental hypothesis being true.
> (5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.
> (6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

Each of the six beliefs is false, a possibility explicitly stated in the instruction. Beliefs 1 and 3 are illusions of certainty: significance tests provide probabilities, not certainties. Beliefs 2, 4, and 5 are versions of Bayesian wishful thinking. Belief 2 is incorrect because a p value is not the probability that the null hypothesis is true but rather the probability of the given data (or more extreme data), assuming the truth of the null hypothesis. For the same reason, one cannot deduce the probability that the experimental (alternative) hypothesis is true, as stated in Belief 4. Belief 5 makes essentially the same claim as Belief 2 because a wrong decision to reject the null hypothesis amounts to the null hypothesis actually being true, and again, the p value does not specify the probability that the null is true. Belief 6 has been dealt with in the previous section. Note that all six delusions

Table 2. Studies on Systematic Delusions About What a Significant *p* Value Means: Percentage of Respondents Who Endorsed the Illusion of Certainty, Bayesian Wishful Thinking, and the Replication Delusion

	Great Britain	Germany			Spain						Italy, Chile	
Delusion	Academic psychologists (N = 70)	Statistics teachers in psychology (N = 30)	Professors of psychology (N = 39)	Students who had passed two statistics courses (N = 44)	Academic psychologists: personality, evaluation, psychological treatments (N = 98)	Academic psychologists: methodology (N = 67)	Academic psychologists: basic psychology (N = 56)	Academic psychologists: social psychology (N = 74)	Academic psychologists: psychobiology (N = 29)	Academic psychologists: developmental and educational psychology (N = 94)	Academic psychologists: methodology (N = 18)	Academic psychologists: other areas (N = 146)
1. The null hypothesis is false.	1	10	15	34	65	36	61	66	55	62	28	33
2. The probability of the null hypothesis being true is known.	36	17	26	32	51	58	68	62	56	58	28	23
3. The alternative hypothesis is true.	6	10	13	20								
4. The probability of the alternative hypothesis being true is known.	66	33	33	59	41	13	23	37	38	44	6	12
5. The probability of Type I error is known.	86	73	67	68								
6. The probability of successful replication is known.	60	37	49	41	35	16	36	39	28	64	6	13
Total (percentage who endorsed at least one delusion)	97	80	90	100	94[a]						56	74

Note: Beliefs 1 and 3 are instances of the illusion of certainty and make successful replication appear certain when the *p* value is significant. Beliefs 2, 4, and 5 are instances of Bayesian wishful thinking and suggest that the *p* value indicates the probability that the null or the alternative hypothesis is true. Belief 6 is the replication delusion (see Table 1); it is included here so that results for all six delusions are summarized in one place. The question that respondents answered referred to $p = .01$ in the studies conducted in Great Britain (Oakes, 1986) and Germany (Haller & Krauss, 2001) and to $p = .001$ in the studies conducted in Spain (Badenes-Ribera, Frias-Navarro, Monterde-i-Bort, & Pascual-Soler, 2015) and Italy and Chile (Badenes-Ribera, Frias-Navarro, Iotti, Bonilla-Campos, & Longobardi, 2016).
[a]Badenes-Ribera et al. (2015) did not report the percentage of respondents who endorsed at least one delusion separately for each subsample in their study, so this is the percentage for all respondents.

err in the same direction of wishful thinking: They overestimate what can be concluded from a p value.

What is most important for the topic of replication is that each of these beliefs makes replication appear to be superfluous. Consider the illusion of certainty (Beliefs 1 and 3): If one concludes that the experimental hypothesis has been absolutely proven to be true or that the null hypothesis has been absolutely disproven, then replicating the study appears to be a waste of time. Similarly, if one incorrectly believes that the p value specifies the probability that the null hypothesis is true (Beliefs 2 and 5) or that the alternative hypothesis is true (Belief 4), then one presumably already knows the Bayesian posterior probability. For instance, if $p = .01$, then someone with these beliefs would misperceive the probability of the alternative hypothesis being true as 99%, which would make further replication attempts appear to be unnecessary. All six delusions lead to a vast overestimation of the probability of successful replications.

Do trained academic psychologists believe in the illusion of certainty and engage in Bayesian wishful thinking? The studies indicate that they do, to varying extents:

- **Belief 1. The null hypothesis has been shown to be false:** Table 2 shows that in every study and subgroup, some professionals held this elementary illusion of certainty. The highest percentages were obtained among Spanish academic psychologists (55%–66%), and the lowest percentage was found among British academic psychologists (1%). Even among professionals teaching statistics or methodology, 10% of German, 28% of Italian and Chilean, and 36% of Spanish experts shared this illusion of certainty.
- **Belief 2. The probability of the null hypothesis being true is known:** Among psychologists who taught methods or statistics, 17% to 58% believed that this conclusion was correct. Among the other professionals, the range was slightly higher, from a minimum of 23% to a maximum of 68%.
- **Belief 3. The alternative hypothesis has been shown to be true:** This second illusion of certainty was included in only a few studies. In all of these cases, it was endorsed by a small percentage of psychologists, including 10% of those who taught statistics.
- **Belief 4. The probability of the alternative hypothesis being true is known:** In every study, some academic psychologists shared this delusion. The percentage who endorsed this belief ranged from 6% to 33% among those who taught methodology and from 12% to 66% among those who did not.
- **Belief 5. The probability of incorrectly rejecting the null hypothesis is known:** Presented in only a few studies, this statement received the highest percentage of endorsements when it was included. The percentage of respondents who agreed with this statement ranged from 67% to 86%, and even a majority (73%) of statistics teachers endorsed it.

Belief 6 (the replication delusion; see Table 1) is included in Table 2 to provide a summary of the results for all six delusions. The last row of the table shows the percentage of respondents who were in the grip of at least one of the delusions: 97% of the British academic psychologists, 80% of the German statistics teachers, 90% of the German psychology professors and lecturers who did not teach statistics, and 100% of the German students who had successfully passed two statistics courses (Table 2, last row). The students appear to have inherited the delusions from their teachers. Among the Spanish academic psychologists, 94% endorsed at least one of the delusions, whereas 56% and 74%, respectively, of the methodology instructors and other academic researchers in Italy and Chile did the same. For the German study, Haller and Krauss (2002) also reported the average number of delusions in each group: 1.9 among statistics teachers, 2.0 among other academic psychologists, and 2.5 among psychology students.

Hoekstra, Morey, Rouder, and Wagenmakers (2014) adapted Oakes's (1986) six-item questionnaire to examine delusions regarding confidence intervals. They reported that the majority of 118 researchers, 34 master's students, and 442 first-year students in psychology relied on similar wishful thinking about confidence intervals; in all three groups, the median of number of delusions endorsed was 3 to 4. Only 3% of the researchers were able to identify all the statements as delusions.

Similarly, Falk and Greenbaum (1995) reported that 87% of 53 Israeli psychology students believed in at least one of the first four delusions listed in Table 2, even though the correct alternative ("None of the answers 1–4 is correct") was added to the response options.

In a study in France, Lecoutre, Poitevineau, and Lecoutre (2003) presented participants with a vignette about a drug that had a significant effect of a small size and found that psychological researchers were more impressed about the efficacy of the drug than were statisticians from pharmaceutical companies. Thus, the

psychological researchers confused statistical significance with substantial significance.

Delusions about significance among medical doctors and researchers

One could argue that delusions about statistical inferences and replicability are embarrassing but largely inconsequential in many areas of psychology: These delusions do not harm the general public's health or wealth. The situation is different in medicine, where manipulating statistics and lack of understanding can lead to death, morbidity, and waste of resources (Welch, 2011). Medical professionals are expected to read medical journals and understand the statistics in order to provide the best treatments to patients.

Do medical doctors and researchers exhibit the same misconceptions, despite these possible adverse consequences? A literature search revealed very few studies on the topic of physicians' understanding of statistical significance. I begin with the one that is closest to those listed in Table 2.

At three major academic U.S. hospitals—the Barnes Jewish Hospital, Brigham & Women's Hospital, and Massachusetts General Hospital—a total of 246 physicians were given the following problem (Westover, Westover, & Bianchi, 2011, p. 1):

> Consider a typical medical research study, for example designed to test the efficacy of a drug, in which a null hypothesis H_0 ('no effect') is tested against an alternative hypothesis H_1 ('some effect'). Suppose that the study results pass a test of statistical significance (that is P-value <0.05) in favor of H_1. What has been shown?
>
> 1. H_0 is false.
> 2. H_0 is probably false.
> 3. H_1 is true.
> 4. H_1 is probably true.
> 5. Both (1) and (3)
> 6. Both (2) and (4)
> 7. None of the above.

Note that the first four statements correspond to the first four beliefs in Table 2, though the wording differs and no precise probability is attached to the null or alternative hypothesis; in addition, a correct answer (7) is offered. Nevertheless, only 6% of the physicians recognized the correct answer, and the remaining 94% believed that a *p* value less than .05 meant that the null hypothesis was false or probably false, or that the alternative hypothesis was true or probably true. Specifically, 4% endorsed the first response option, 31% endorsed the second, none endorsed the third, 20% endorsed the fourth, 3% endorsed the fifth, and 36% endorsed the sixth.

B. L. Anderson, Williams, and Schulkin (2013) tested U.S. obstetrics-gynecology residents (i.e., beginning doctors) participating in the Council for Resident Education in Obstetrics and Gynecology In-Training Examination. Obstetrics-gynecology is a prestigious specialty that attracts students with very good grades in medical school. The response rate to the survey was 95% (4,713 out of 4,961 residents). The delusion question Anderson et al. presented to the residents was similar to Belief 2 in Table 2: "True or False: The P value is the probability that the null hypothesis is correct" (p. 273). Forty-two percent of the respondents correctly answered "false," 12% did not answer, and 46% incorrectly said "true." Nevertheless, 63% of the respondents rated their statistical literacy as adequate, 8% rated it as excellent, and only 22% rated it as inadequate (7% did not respond).

Wulff, Andersen, Brandenhoff, and Guttler (1987) tested 148 Danish doctors (randomly sampled) and 97 participants in a postgraduate course in research methods, mainly junior hospital doctors. When asked what it means if a controlled trial shows that a new treatment is significantly better than placebo ($p < .05$), 20% of the doctors in the random sample and 6% of the participants in the postgraduate course) said that "it has been proved that the treatment is better than placebo"; 51% and 54%, respectively, believed that the probability of the null hypothesis being true is less than .05 (Bayesian wishful thinking); and 18% and 2%, respectively, said that they did not know what *p* values mean. Only 13% of the doctors in the random sample and 39% of the participants in the course could identify the correct answer ("If the treatment is not effective, there is less than a 5 per cent chance of obtaining such results," p. 5).

These few available studies suggest that in medicine, where decisions are consequential, the same delusions regarding the null ritual appear to persist. This interpretation is supported by other tests of physicians' statistical literacy (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Wegwarth, Schwartz, Woloshin, Gaissmaier, & Gigerenzer, 2012). The systematic errors "have been encouraged if not licensed by unjustified, lax, or erroneous traditions and training in the field at large" (Greenland, 2011, p. 228).

Discussion

Taken together, the studies indicate that a substantial proportion of academic researchers wrongly believe that the *p* value obtained in a study implies that the probability of finding another significant effect in a

replication is 1 – p. In the view of researchers who share the other delusions I have been discussing, replication studies are superfluous because the p value already provides certainty or at least high probability of a successful replication. Thus, for these researchers, a failed replication may come as a total surprise.

If a problem of incentives only were at the heart of the replication crisis, these delusions would not likely exist. Conversely, the thesis that many researchers have internalized a statistical ritual implies the existence of delusions that maintain the ritual.

The results in Tables 1 and 2 are supported by studies of how significant results are interpreted in published articles. Finch, Cumming, and Thomason (2001) reviewed the articles published in the *Journal of Applied Psychology* over 60 years and concluded that in 38% of these articles, nonsignificance was interpreted as demonstrating that the null hypothesis was true (illusion of certainty). Similarly, an analysis of 259 articles in the *Psychonomic Bulletin & Review* revealed that in 19% of the articles, the authors presented statistical significance as certainty (Hoekstra et al., 2006). These values are consistent with those in Table 2.

Statistical power and replication

In order to investigate whether an effect exists, one should design an experiment that has a reasonable chance of detecting it. I take this insight as common sense. In statistical language, an experiment should have sufficient statistical power.

Yet the null ritual knows no statistical power. Early textbook writers such as Guilford declared that the concept of power was "too difficult to discuss" (1956, p. 217). The fourth edition of the *Publication Manual of the American Psychological Association* was the first to mention that power should be taken seriously (American Psychological Association, 1994), but no practical guidelines were given. Nor did the subsequent fifth and sixth editions provide guidelines (American Psychological Association, 2001, 2010), which is odd in a manual that instructs authors on minute details such as how to format variables and when to use a semicolon.

Statistical power is a concept from Neyman-Pearson theory: Power is the probability of accepting the alternative hypothesis if it is true. For instance, if the alternative hypothesis is true and the power is 90%, and if the experiment is repeated many times, one will correctly conclude in 90% of the cases that this hypothesis is true. Power is directly relevant for the probability of a successful replication in two respects. First, if the alternative hypothesis is correct but the power is low, then the chances of replicating a significant finding are low. Second, if the power is low, then significant findings overestimate the size of the effect, which is one of the reasons why effects—even those that exist—tend to "fade away" (Button et al., 2013).

Statistical power provides another test of whether the incentive structure of "publish or perish" is sufficient to explain the replication crisis or whether a substantial part of this crisis is due to the null ritual and its associated delusions. Given the incentive structure for producing statistically significant results, it should be in the interest of every researcher to design experiments that have a reasonable chance of detecting an effect. Thus, according to the strategic-game hypothesis, we would expect researchers to strategically design experiments with high or at least reasonable power (with the exception of experiments for which the effect size is expected to be small; see the Discussion section). A minimal criterion for "reasonable" would be "substantially better than a coin toss." That is, if one performs a chance experiment by tossing a coin and accepts the alternative hypothesis if "heads" comes up, this "experiment" has power of 50% to correctly "detect" an effect if there is one. Any psychological experiment should be designed to have a better power. However, to the degree that researchers have internalized the null ritual, which does not know power, we would expect both inattention to power, which results in small power, and unawareness of this problem. Thus, the statistical-ritual hypothesis predicts that researchers act against their own best interest.

Better than a coin flip? Cohen (1962) estimated the power of studies published in the *Journal of Abnormal and Social Psychology* for detecting what he called small, medium, and large effect sizes (corresponding to Pearson correlations of .2, .4, and .6, respectively). He reported that the median power to detect a medium-sized effect was only 46%. A quarter of a century later, Sedlmeier and I checked whether Cohen's study on power had had an effect on the power of studies in the *Journal of Abnormal Psychology* (Sedlmeier & Gigerenzer, 1989). It had not; the median power to detect a medium-sized effect had decreased to 37%.[3] The decline was a result of the introduction of alpha-adjustment procedures, reflecting the focus of the null ritual on the p value. Low power appeared to go unnoticed: Only 2 of 64 reports mentioned power at all. Subsequently, we checked the years 2000 through 2002 of the same journal and found that just 9 out of 220 empirical articles included statements about how the researchers determined the power of their tests (Gigerenzer, Krauss, & Vitouch, 2004). Bakker, Hartgerink, Wicherts, and van der Maas (2016) reported that 89% of 214 authors overestimated the power of research designs. Other analyses showed that only 3% of 271 psychological articles reporting significance tests

explicitly discussed power as a consideration for designing an experiment (Bakker & Wicherts, 2011), and only 2% of 436 articles in the *Journal of Speech, Language, and Hearing Research* in 2009 through 2012 reported statistical power (Rami, 2014). A meta-analysis of 44 reviews published in the social and behavioral sciences, beginning with Cohen's 1962 study, found that power had not increased over half a century (Smaldino & McElreath, 2016). Instead, the average power had remained consistently low, and the mean power for detecting a small-sized effect (Cohen's $d = 0.2$) was 24%, assuming $\alpha = .05$. An analysis of 3,801 cognitive neuroscience and psychology articles published between 2011 and 2014 found that the median power to detect small, medium, and large effects was 12%, 44%, and 73%, respectively; in other words, there had been no improvement since the first power studies were conducted (Szucs & Ioannidis, 2017).

To estimate the power of experiments, an alternative route is to estimate the main factors that affect power: sample size and effect size. Given the median total sample size of 40 in four representative journals (*Journal of Abnormal Psychology, Journal of Applied Psychology, Journal of Experimental Psychology: Human Perception and Performance*, and *Developmental Psychology*; Marszalek, Barber, Kohlhart, & Holmes, 2011) and the average meta-analytic effect size (d) of 0.50, Bakker et al. (2012) estimated that the typical power of psychology studies is around 35%.

In the neurosciences, power appears to be exceptionally low. An analysis of meta-analyses that included 730 individual neuroscience studies—on the genetics of Alzheimer's disease, brain-volume abnormalities, cancer biomarkers, and other topics—revealed that the median statistical power of the individual studies to be able to detect an effect of the summary effect size reported in their corresponding meta-analyses was 21% (Button et al., 2013). The distribution of power was bimodal. A few, mostly with a neurological focus, had a power greater than 90%, whereas the power of the 461 structural and volumetric MRI studies was strikingly low, 8%. Among the animal-model studies included in this analysis, the average power was 20% to 30%. This suggests the possibility that animals' lives are typically sacrificed in poorly designed experiments that have low chances of finding an effect. Simply flipping a coin would be a better strategy, sparing both the animals and the resources involved.

Discussion. If statistical significance is so devoutly desired by behavioral scientists, why do they design experiments that typically have such low chances of finding a significant result? Consistently low power over some 50 years is not expected under the hypothesis that researchers strategically aim at achieving statistically significant results. But it is consistent with the hypothesis that they are following the null ritual, which knows no power.

However, there may be a strategic element in designing low-power studies if the expected effect size is small and one runs multiple studies. In this situation, the chance of at least one significant result in N low-powered studies with sample size n/N can be higher than the chance of a significant result in one high-powered study with sample size n (Bakker et al., 2012). Although it is unlikely that most researchers strategically reason this way, given the general lack of thinking about power, positive experience could well reinforce such behavior. There is also the possibility that researchers sometimes design a single low-powered study so as to engineer nonsignificance, such as when they want to "demonstrate" the absence of adverse side effects of drugs (Greenland, 2012). Thus, the typical lack of statistical power is implied by the statistical-ritual hypothesis, but it may also have a strategic element.

At the same time, analyses showing that studies are frequently low powered raise a new question. Why do more than 90% of published articles in major psychological journals report significant results, despite notoriously low power (Sterling, Rosenbaum, & Weinkam, 1995)? The answer appears to be that many researchers compensate for their blind spot regarding power by violating good scientific practice in order to nevertheless produce significant results. In one study, 2,155 academic psychologists at major U.S. universities agreed to report anonymously whether they had personally engaged in questionable research practices; half of the psychologists received incentives to answer honestly (John, Loewenstein, & Prelec, 2012). To control for reporting bias and estimate the true prevalence, the researchers also asked all the psychologists to estimate the percentage of other psychologists who had engaged in the same questionable behaviors and, among those who had, the percentage who would actually admit to having done so. For the group with incentives to report honestly, the seven most frequent questionable practices were as follows (each practice is followed by the percentage of the group who admitted to engaging in it and, in parentheses, the estimated true prevalence):

1. Failing to report all dependent measures: 67% (78%)
2. Collecting more data after seeing whether results were significant: 58% (72%)
3. Selectively reporting studies that "worked": 50% (67%)
4. Excluding data after looking at the impact of doing so on the results: 43% (62%)
5. Reporting an unexpected finding as having been predicted from the start: 35% (54%)

6. Failing to report all of a study's conditions: 27% (42%)
7. Rounding down a *p* value (e.g., reporting .054 as less than .05): 23% (39%)

By violating the statistical model on which the *p* value depends, each of these practices makes a significant result noninterpretable. The first practice increases the chance of a significant result from the nominal 5% if the null hypothesis is correct to a larger value, depending on the number of dependent variables (Simmons, Nelson, & Simonsohn, 2011). Among the 2,155 psychologists, 67% admitted that they had not reported all the measures they had used, and the estimated true value was higher. The other practices serve the same goal: to inflate the production of significant results. The last practice, rounding down *p* values to make them appear significant, is clearly cheating. It can be independently uncovered because it produces a systematic gap in the distribution of *p* values: too few just above .05 and too many just under .05. This pattern was found, for instance, in reports concluding that food ingredients cause cancer (Schoenfeld & Ioannidis, 2013). The practice of rounding down *p* values can also be detected from inconsistencies between reported *p* values and reported test statistics. Among articles published in 2001 in the *British Medical Journal* and in *Nature*, 25% and 38%, respectively, reported that results were statistically significant even though the test statistics revealed that they were not (García-Berthou & Alcaraz, 2004). Among the psychologists in the study by John et al. (2012), a similar percentage admitted to rounding down *p* values. The R package statcheck can help detect inconsistent *p* values (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016).

The percentages reported by John et al. (2012) are likely conservative, given that out of some 6,000 psychologists originally contacted by the authors, only 36% responded. The low response rate could reflect a self-selection bias that resulted in more honest researchers being more likely to participate in the survey. Despite this possibility, a total of 94% of the researchers admitted that they had engaged in at least one questionable research practice.

General Discussion

The argument

I have argued that the replication crisis in psychology and the biomedical sciences is not only a matter of wrong incentives that are gamed by researchers (the strategic-game hypothesis) but also a consequence of researchers' belief in the null ritual and its associated delusions (the statistical-ritual hypothesis). In the first section of this article, I reconstructed the creation of the "null ritual" by textbook writers who merged two competing statistical theories into one hybrid theory, whose core is the null ritual and whose desired product is statistical significance. This ritual eventually replaced good standards of scientific practice with a single convenient surrogate: the *p* value.

In the second section, I tested four predictions of the statistical-ritual hypothesis. The first of these is that a substantial proportion of academic researchers should share the replication delusion. A review of the available studies with 839 academic psychologists and 991 psychology students showed that 20% of the faculty teaching statistics in psychology, 39% of the professors and lecturers, and 66% of the students did so.

The second and third predictions are that a substantial proportion of researchers should share the illusion of certainty and Bayesian wishful thinking, respectively. In the studies I reviewed, between 56% and 80% of statistics and methodology instructors in psychology departments believed in one or more of these three delusions; this range increased to 74% to 97% for professors and lecturers who were not methodology specialists. To see through these delusions does not require understanding of high-level statistics; in other contexts, researchers themselves study whether their participants are subject to the illusion of certainty or the inverse probability error (e.g., Hafenbrädl & Hoffrage, 2015).

The fourth prediction of the statistical-ritual hypothesis is that researchers should be largely blind to statistical power because it is not part of the ritual. The available meta-analyses in psychology show that the median power to detect a medium-sized effect is around 50% or below, which amounts to the power of tossing a coin. There has been no noticeable improvement since the first power analysis in the 1960s.

The statistical-ritual hypothesis also explains why an estimated 94% of academic psychologists engage in questionable research practices to obtain significant results (John et al., 2012). Significance, that is, rejection of the null hypothesis, is the primary goal of the null ritual, relegating good scientific practice to a secondary role. Researchers do not engage in questionable practices to minimize measurement error or to derive precise predictions from competitive theories; they engage in these practices solely in order to achieve statistical significance.

What to do

Various proposals have been made to prevent questionable practices by changing the incentive structure and introducing measures such as preregistration of studies (e.g., Gigerenzer & Muir Gray, 2011; Ioannidis, 2014). Trust in science could be improved by preregistration,

but given the fixation on significance, this important measure has been already gamed: In medicine, systematic discrepancies between registered and published experiments are common, registration often occurs after the data have been obtained, reviewers do not take the time to compare the registered protocol with the submitted article, and, despite preregistration, selective reporting of outcomes to achieve significance is frequent (C. W. Jones, Keil, Holland, Caughey, & Platts-Mills, 2015; Walker, Stevenson, & Thornton, 2014). As an alternative measure, a group of 72 researchers proposed redefining the criterion for statistical significance as $p < .005$ rather than $p < .05$ (Benjamin et al., 2017). The authors concluded: "The new significance threshold will help researchers and readers to understand and communicate evidence more accurately" (p. 11). Although this measure would be useful for reducing false positives, I do not see how it would improve understanding and eradicate the delusions documented in Tables 1 and 2. For researchers who believe in the replication fallacy, a p value less than .005 means that the results can be replicated with a probability of 99.5%.

I now provide a complementary proposal that follows from the present analysis. This proposal consists of four steps, the first of which would serve as a minimal solution by eliminating the surrogate goal. The second through fourth steps would extend this solution by refocusing research on good scientific method. The ultimate goal of this proposal is to support statistical thinking instead of statistical rituals.

Step 1: editors should no longer accept manuscripts that report results as "significant" or "not significant." If p values are reported, they should be reported as exact p values—for example, $p = .04$ or .06—as are other continuous measures, such as effect sizes. Decisions about accepting or rejecting an article should be based solely on its theoretical and methodological qualities, regardless of the p values.

This measure would eliminate the surrogate goal of reaching "significance" and the associated pressure to sacrifice proper scientific method in order to get a significant result. Science is a cumulative endeavor, not a yes/no decision based on a single empirical study. This step is a minimal proposal because it is easy to implement, but more radical than the alternative proposal to lower the level of significance (Benjamin et al., 2017). Lowering p values does not eliminate the surrogate goal but only makes it more difficult to attain. Although this increased difficulty might make some forms of p-hacking less effective, it may encourage even more concentration on the p value and increase questionable research practices used to attain significant results, thereby diverting attention from good scientific practice.

Step 2: editors should make a distinction between research aimed at developing hypotheses and research aimed at testing hypotheses. Editors should require that researchers clearly distinguish between developing hypotheses (e.g., looking through a correlation matrix to find large correlations) and testing hypotheses (e.g., running a second experiment in which this large correlation is stated as a hypothesis and subsequently tested). This distinction is also known as the distinction between exploratory and confirmatory research (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). In order to make the distinction transparent, authors should not report any p values or similar inferential statistics when they report research aimed at developing hypotheses. Editors should encourage researchers to report on both hypothesis development and hypothesis testing.

A clear distinction between these two types of research would encourage direct replication attempts (i.e., independent tests of interesting observations made in prior experiments). It would also relieve researchers from the pressure to present unexpected findings as having been predicted from the start (see John et al., 2012).

Step 3: editors should require competitive-hypothesis testing, not null-hypothesis testing. Editors should require that a new research hypothesis is tested against the best competitors available. Null-hypothesis testing, in contrast, is noncompetitive: Typically, the prediction of the research hypothesis remains unspecified and is tested against a null effect only.

Competitive testing requires precise research hypotheses and thus encourages building mathematical models of psychological processes. Competition would make null-hypothesis testing obsolete. It could also improve the impoverished approach to theory in psychology, which is partly due to the focus on null-hypothesis testing (Gigerenzer, 1998). If the competing models use free parameters, these need to be tested in prediction (e.g., out-of-sample prediction, such as cross-validation), never by data fitting alone (e.g., Brandstätter, Gigerenzer, & Hertwig, 2008).

Step 4: psychology departments should teach the statistical toolbox, not a statistical ritual. Psychology departments need to begin teaching the statistical toolbox. This toolbox includes techniques to visualize the descriptive statistics of the sample, Tukey's exploratory data analysis, meta-analysis, estimation, Fisher's null-hypothesis testing (which is not the same as the null ritual), Neyman-Pearson decision theory, and Bayesian inference. Most important, the toolbox approach requires using informed judgment to select the appropriate tool for a given problem.

The emphasis on judgment would mean taking the assumptions of statistical models seriously. Editors should not require routine statistical inference in situations in which it is unclear whether the assumptions of a statistical model actually hold but rather should encourage proper use of descriptive statistics or exploratory data analysis. The toolbox approach replaces statistical rituals with statistical thinking and includes principles of good scientific method, such as minimizing measurement error and conducting double-blind studies.

The key challenge in the toolbox approach is to develop informed judgment about the kind of problems that each tool can handle best—a process similar to learning that hammers are for nails and screwdrivers are for screws. Fisher's null-hypothesis testing is useful (if at all) solely for new problems for which little information is available and one does not even have a precise alternative hypothesis. If two competing hypotheses are known, Neyman-Pearson theory is the better choice. If, in addition, priors are known, as in cancer screening, Bayes rule is likely the preferred method; of the three tools, it is also the only one designed to estimate probabilities that hypotheses are true. Most important, however, high-quality descriptive statistics and exploratory data analysis (Breiman, 2001; L. V. Jones & Tukey, 2000) are good candidates for the scores of situations in which no random samples have been drawn from defined populations and the assumptions of the statistical models are not in place (Greenland, 1990). Such a toolbox approach is the opposite of an automatic inference procedure. It requires good judgment about when to use each tool, which is exactly what Fisher, Neyman and Pearson, and Jones and Tukey emphasized.

The toolbox approach can correct the historical error of considering statistical inference from sample to population as the sine qua non of good scientific practice. This has been an extraordinary blunder, and for two reasons. First, as mentioned before, the assumptions underlying the model of statistical inference are typically not met. For instance, typically no population has been defined, and no random samples have been drawn. Thus, unlike in quality control or polling, nobody knows the population to which a significant result actually refers, which makes the entire *p*-value procedure a nebulous exercise. Second, and most important, obtaining statistical significance has become a surrogate for good scientific practice, pushing principles such as formulating precise theories, conducting double-blind experiments, minimizing measurement error, and replicating findings into the sidelines. These principles have often not even been mentioned in reports on phenomena that later proved difficult to replicate, including reports on priming and too-much-choice experiments. W. S. Gosset, who published an article on the *t* test in 1908, recognized this long ago with respect to measurement error: "Obviously the important thing . . . is to have a low real error, not to have a 'significant' result at a particular station [level]. The latter seems to me to be nearly valueless in itself" (quoted in Pearson, 1939, p. 247).

Incentives

Finally, the established incentives themselves need explanation. Why do they arbitrarily focus on statistical significance, which by itself is one of the least important signs of good scientific research? In theory, researchers could be rewarded for quite a number of practices, including demonstrating the stability of effects through replications and designing clever experiments that discriminate between two or more competing hypotheses. For instance, physicists are rewarded for designing tools that minimize measurement error, and neoclassical economists are rewarded for developing mathematical models, theorems, and proofs. Whereas the strategic-game hypothesis takes the incentives as given, the statistical-ritual hypothesis provides a deeper explanation of the roots of the replication crisis. Researchers are incentivized to aim for the product of the null ritual, statistical significance, not for goals that are ignored by it, such as high power, replication, and precise competing theories and proofs. The statistical-ritual hypothesis provides the rationale for the very incentives chosen by editors, administrators, and committees. Obtaining significant results became the surrogate for good science.

Surrogate science: the elimination of scientific judgment

The null ritual can be seen as an instance of a broader movement toward replacing judgment about the quality of research with quantitative surrogates. Search committees increasingly tend to look at applicants' h-indices, citation counts, and numbers of articles published as opposed to actually reading and judging the arguments and evidence provided in these articles. Assessment of research is also increasingly left to administrators who do not understand the content of the research, a practice encouraged by the rising commercialization of universities and of academic publishing.

One positive aspect of the replication crisis is that it has increased awareness that we need to take action and protect science from being transformed into a mass production of studies that pursue surrogate goals. What we need are not more but fewer and better publications. In order to ensure that future generations of scientists remain innovative risk takers, educators, journal editors, and researchers themselves need to revert

to the original goals of science. It is time to combat the present system of false incentives and eliminate the null ritual from scientific practice.

Action Editor

Daniel J. Simons served as action editor for this article.

Author Contributions

G. Gigerenzer is the sole author of this article and is responsible for its content.

Acknowledgments

I would like to thank the ABC Research Group, Henry Cowles, Geoff Cumming, Sander Greenland, Deborah Majo, Rona Unrau, and E.-J. Wagenmakers for their helpful comments. This article is based on a seminar I presented at the Davis Center, History Department, Princeton University, October 2016, and a talk I gave at the Philosophy of Science meeting, Atlanta, Georgia, November 2016.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Notes

1. Reporting replication results in yes/no terms can be just as problematic as reporting significance in yes/no terms. If an original conclusion that an effect exists was based on statistical significance alone and the replicability of the effect is again determined by significance, then the flaws of relying on *p* values alone carry over to the replication studies.

2. During his career, Fisher changed his ideas on this and other questions, in part motivated by his controversy with Neyman and Pearson. Earlier, he had proposed .05 as a convenient level of significance, but in the 1950s he rejected the routine use of such a constant level. Thus, Fisher himself may have contributed to the confusion. Similarly, from reading his *Design of Experiments* (Fisher, 1935), one might gain the impression that null-hypothesis testing is fairly mechanical, but Fisher later made it quite clear that this was not his intention (see Gigerenzer et al., 1989, chap. 3).

3. In this replication study, we used Cohen's original definition of a medium effect size to facilitate comparison with his original study, although Cohen (1969) later changed his definition of small, medium, and large effect sizes to correspond to Pearson correlations of .1, .3, and .5, respectively. This systematic lowering of the effect-size convention has the effect of slightly lowering the power, too. These assumed effect sizes may still be larger than those typical in some fields. In applied psychology, tertile effect sizes (calculated by dividing the distribution of effect sizes into three equal parts) have been reported to be only half or a third as large as Cohen's revised values (Bosco, Aguinis, Singh, Field, & Pierce, 2015). In what follows, I report power estimates as published without discussing the details, which would go beyond the focus of this article (for a discussion of counterintuitive issues involving power, see Greenland, 2012). But a word of caution is necessary. Not all estimated power values can be directly compared because they may be based on differing assumptions.

References

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Anastasi, A. (1958). *Differential psychology* (3rd ed.). New York, NY: Macmillan.

Anderson, B. L., Williams, S., & Schulkin, J. (2013). Statistical literacy of obstetrics-gynecology residents. *Journal of Graduate Medical Education*, *5*, 272–275. doi:10.4300/JGME-D-12-00161.1

Anderson, R. L., & Bancroft, T. A. (1952). *Statistical theory in research*. New York, NY: McGraw-Hill.

Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A., & Longobardi, C. (2016). Misconceptions of the *p*-value among Chilean and Italian academic psychologists. *Frontiers in Psychology*, *7*, Article 1247. doi:10.3389/fpsyg.2016.01247

Badenes-Ribera, L., Frias-Navarro, D., Monterde-i-Bort, H., & Pascual-Soler, M. (2015). Interpretation of the p value: A national survey study in academic psychologists from Spain. *Psicothema*, *27*, 290–295. doi:10.7334/psicothema2014.283

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*, 1069–1077. doi:10.1177/0956797616647519

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. doi:10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666–678. doi:10.3758/s13428-011-0089-5

Benjamin, D. J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. (2017). *Redefine statistical significance*. Retrieved from psyarxiv.com/mky9j

Bokhari, A. (2017, March 29). J Scott Armstrong: Fewer than 1 percent of papers in scientific journals follow scientific method. *breitbart.com*. Retrieved from http://www.breitbart.com/tech/2017/03/29/j-scott-armstrong-fraction-1-papers-scientific-journals-follow-scientific-method/

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*, 431–449. doi:10.1037/a0038047

Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist*, *27*, 774–775. doi:10.1037/h0038034

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, *113*, 409–432. doi:10.1037/0033-295X.113.2.409

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*, 199–231. doi:10.1214/ss/1009213726

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376. doi:10.1038/nrn3475

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153. doi:10.1037/h0045186

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of *p*-values. *Royal Society Open Science*, *1*(3), Article 140216. doi:10.1098/rsos.140216

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286–300. doi:10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. doi:10.1177/0956797613504966

Danziger, K. (1987). Statistical method and the historical development of research practice in American psychology. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution, Vol. 2: Ideas in the sciences* (pp. 35–47). Cambridge, MA: MIT Press.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. Cambridge, England: Cambridge University Press.

Dulaney, S., & Fiske, A. P. (1994). Cultural rituals and obsessive-compulsive disorder: Is there a common psychological mechanism? *Ethos*, *22*, 243–283. doi:10.1525/eth.1994.22.3.02a00010

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard. *Theory & Psychology*, *5*, 75–98. doi:10.1177/0959354395051004

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement*, *61*, 181–210. doi:10.1177/00131640121971167

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B*, *17*, 69–78.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh, Scotland: Oliver & Boyd.

Freedman, L. P., Cockburn, I. A., & Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLOS Biology*, *13*(6), Article e1002165. doi:10.1371/journal.pbio.1002165

García-Berthou, E., & Alcaraz, C. (2004). Incongruence between test statistics and *P* values in medical papers. *BMC Medical Research Methodology*, *4*, Article 13. doi:10.1186/1471–2288-4-13

Garcia-Pérez, M. A., & Alcalá-Quintana, R. (2016). The interpretations of scholars' interpretations of confidence intervals: Criticism, replication, and extension of Hoekstra et al. (2014). *Frontiers in Psychology*, *7*, Article 1042. doi:10.3389/fpsyg.2016.01042

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution, Vol. 2. Ideas in the sciences* (pp. 11–33). Cambridge, MA: MIT Press.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 313–339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology*, *8*, 195–204. doi:10.1177/0959354398082006

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*, 587–606. doi:10.1016/j.socec.2004.09.033

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients to make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96. doi:10.1111/j.1539-6053.2008.00033.x

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about null hypothesis testing but were afraid to ask. In D. Kaplan (Ed.), *Handbook on quantitative methods in the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Gigerenzer, G., & Muir Gray, J. A. (Eds.). (2011). *Better doctors, better patients, better decisions: Envisioning health care 2020*. Cambridge, MA: MIT Press.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge, England: Cambridge University Press.

Greenland, S. (1990). Randomization, statistics, and causal inference. *Epidemiology*, *1*, 421–429.

Greenland, S. (2011). Null misinterpretation in statistical testing and its impact on health risk assessment. *Preventive Medicine*, *53*, 225–228.

Greenland, S. (2012). Nonsignificance plus high power does not imply support for the null over the alternative. *Annals of Epidemiology*, *22*, 364–368. doi:10.1016/j.annepidem.2012.02.007

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology*, *33*, 175–183. doi:10.1111/j.1469-8986.1996.tb02121.x

Guilford, J. P. (1942). *Fundamental statistics in psychology and education* (1st ed.). New York, NY: McGraw-Hill.

Guilford, J. P. (1956). *Fundamental statistics in psychology and education* (3rd ed.). New York, NY: McGraw-Hill.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education* (4th ed.). New York, NY: McGraw-Hill.

Hafenbrädl, S., & Hoffrage, U. (2015). Toward an ecological analysis of Bayesian inference: How task characteristics influence responses. *Frontiers in Psychology*, *6*, Article 939. doi:10.3389/fpsyg.2015.00939

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online*, *7*(1), 1–20. Retrieved from https://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf

Hoekstra, H., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of *p* values. *Psychonomic Bulletin & Review*, *13*, 1033–1037. doi:10.3758/BF03213921

Hoekstra, H., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 1157–1164. doi:10.3758/s13423-013-0572-3

Horton, R. (2016). Offline: What is medicine's 5 sigma? *The Lancet*, *385*, 1380.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), Article e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2014). How to make more published research true. *PLOS Medicine*, *11*(10), Article e1001747. doi:10.1371/journal.pmed.1001747

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., . . . Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet*, *383*, 166–175. doi:10.1016/s0140-6736(13)62227-8

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:10.1177/0956797611430953

Jones, C. W., Keil, L. G., Holland, W. C., Caughey, M. C., & Platts-Mills, T. F. (2015). Comparison of registered and published outcomes in randomized controlled trials: A systematic review. *BMC Medicine*, *13*, Article 282. doi:10.1186/s12916-015-0520-3

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, *5*, 411–414. doi:10.1037/1082-989X.5.4.411

Lecoutre, M. P., Poitevineau, J., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology*, *38*, 37–45. doi:10.1080/00207590244000250

Lehrer, J. (2010, December 13). The truth wears off: Is there something wrong with the scientific method? *The New Yorker*. Retrieved from https://www.newyorker.com/magazine/2010/12/13/the-truthwears-off

Loftus, G. R. (1993). Editorial comment. *Memory & Cognition*, *21*, 1–3. doi:10.3758/BF03211158

Madden, C. S., Easley, R. W., & Dunn, M. G. (1995). How journal editors view replication research. *Journal of Advertising*, *24*, 77–87.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*, 537–542. doi:10.1177/1745691612460688

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*, 331–348. doi:10.2466/03.11.pms.112.2.331-348

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology*, *64*, 553–557. doi:10.1037/h0045549

Mirowski, P. (2011). *Science-mart: Privatizing American science*. Cambridge, MA: Harvard University Press.

Mullard, A. (2011). Reliability of 'new drug target' claims called into question. *Nature Reviews Drug Discovery*, *10*, 643–644. doi:10.1038/nrd3545

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, *231*, 289–337. doi:10.1098/rsta.1933.0009

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205–1226. doi:10.3758/s13428-015-0664-2

Nunnally, J. C. (1975). *Introduction to statistics for psychology and education*. New York, NY: McGraw-Hill.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York, NY: Wiley.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, Article aac4716. doi:10.1126/science.aac4716

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments. *PLOS ONE*, *7*(8), Article e42510. doi:10.1371/journal.pone.0042510

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. doi:10.1177/1745691612465253

Pearson, E. S. (1939). "Student" as statistician. *Biometrika*, *30*, 210–250. doi:10.2307/2332648

Pearson, E. S. (1962). Some thoughts on statistical inference. *Annals of Mathematical Statistics*, *33*, 394–403. doi:10.1214/aoms/1177704566

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*, 712. doi:10.1038/nrd3439-c1

Rami, M. K. (2014). Power and effect size measures: A census of articles published from 2009-2012 in the *Journal of Speech, Language, and Hearing Research*. *American International Journal of Social Science*, *3*, 13–19.

Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519–559). Hillsdale, NJ: Erlbaum.

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review

of choice overload. *Journal of Consumer Research*, *37*, 409–425. doi:10.1086/651235

Schoenfeld, J. D., & Ioannidis, J. P. A. (2013). Is everything we eat associated with cancer? A systematic cookbook review. *American Journal of Clinical Nutrition*, *97*, 127–134. doi:10.3945/ajcn.112.047142

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*, 437. doi:10.1038/470437

Schünemann, H., Ghersi, D., Kreis, J., Antes, G., & Bousquet, J. (2011). Reporting of research: Are we in for better health care by 2020? In G. Gigerenzer & J. A. Muir Gray (Eds.), *Better doctors, better patients, better decisions: Envisioning health care 2020* (pp. 83–102). Cambridge, MA: MIT Press.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316. doi:10.1037/0033-2909.105.2.309

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), Article 160384. doi:10.1098/rsos.160384

Snedecor, G. W. (1937). *Statistical methods* (1st ed.). Ames: Iowa State Press.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*, 30–34. doi:10.2307/2282137

Sterling, T. D., Rosenbaum, W., & Weinkam, J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*, 108–112. doi:10.2307/2684823

Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59–71. doi:10.1177/1745691613514450

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), Article e2000797. doi:10.1371/journal.pbio.2000797

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110. doi:10.1037/h0031322

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638. doi:10.1177/1745691612463078

Walker, K. F., Stevenson, G., & Thornton, J. G. (2014). Discrepancies between registration and publication of randomised controlled trials: An observational study. *Journal of the Royal Society of Medicine Open*, *5*(5). doi:10.1177/2042533313517688

Wegwarth, O., Schwartz, L. M., Woloshin, S., Gaissmaier, W., & Gigerenzer, G. (2012). Do physicians understand cancer screening statistics? A national survey of primary care physicians in the United States. *Annals of Internal Medicine*, *156*, 340–349. doi:10.7326/0003-4819-156-5-201203060-00005

Welch, H. G. (2011). *Overdiagnosed: Making people sick in the pursuit of health*. Boston, MA: Beacon Press.

Westover, M. B., Westover, K. D., & Bianchi, M. T. (2011). Significance testing as perverse probabilistic reasoning. *BMC Medicine*, *9*, Article 20. doi:10.1186/1741-7015-9-20

Wulff, H. R., Andersen, B., Brandenhoff, P., & Guttler, F. (1987). What do doctors know about statistics? *Statistics in Medicine*, *6*, 3–10.